Skeletal Radiology (2019) 48:239–244
https://doi.org/10.1007/s00256-018-3016-3

SCIENTIFIC ARTICLE

Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network

Takaaki Urakawa[1,2] · Yuki Tanaka[1] · Shinichi Goto[1] · Hitoshi Matsuzawa[3] · Kei Watanabe[2] · Naoto Endo[2]

Received: 15 February 2018 / Revised: 15 June 2018 / Accepted: 18 June 2018 / Published online: 28 June 2018

Abstract

Objective To compare performances in diagnosing intertrochanteric hip fractures from proximal femoral radiographs between a convolutional neural network and orthopedic surgeons.

Materials and methods In total, 1773 patients were enrolled in this study. Hip plain radiographs from these patients were cropped to display only proximal fractured and non-fractured femurs. Images showing pseudarthrosis after femoral neck fracture and those showing artificial objects were excluded. This yielded a total of 3346 hip images (1773 fractured and 1573 non-fractured hip images) that were used to compare performances between the convolutional neural network and five orthopedic surgeons.

Results The convolutional neural network and orthopedic surgeons had accuracies of 95.5% (95% CI = 93.1–97.6) and 92.2% (95% CI = 89.2–94.9), sensitivities of 93.9% (95% CI = 90.1–97.1) and 88.3% (95% CI = 83.3–92.8), and specificities of 97.4% (95% CI = 94.5–99.4) and 96.8% (95% CI = 95.1–98.4), respectively.

Conclusions The performance of the convolutional neural network exceeded that of orthopedic surgeons in detecting intertrochanteric hip fractures from proximal femoral radiographs under limited conditions. The convolutional neural network has a significant potential to be a useful tool for screening for fractures on plain radiographs, especially in the emergency room, where orthopedic surgeons are not readily available.

Keywords Fracture · Deep learning · Orthopedics · Convolutional neural network

Introduction

Artificial intelligence (AI) has been used for several different types of orthopedic diagnoses [1–4]. Olczak et al. detected fractures from plain radiographs using convolutional neural networks (CNNs) [3], which are frequently used AI systems for classifying medical images. Their report indicated that the best performance was achieved using the Visual Geometry Group 16-layer (VGG_16) network [5], one of the CNNs, and the result had an accuracy of 83% for detecting fractures from plain radiographs of the hands, wrists, and ankles. The accuracy was comparable to that of diagnoses made by radiologists. Subsequently, CNNs were applied for wrist fracture detection [6] and proximal humerus fracture detection [7], and they achieved superior network accuracy (96%) compared to that of orthopedic surgeons (93%) [7].

The number of patients with hip fractures is still drastically increasing in Japan [8]. When a hip fracture is suspected based on findings from a physical examination, anterior-view and lateral-view plain radiographs are usually initially taken. However, 2.7% of patients with hip fractures were reported to have unclear fractures on initial plain radiographs, and they required further examination by magnetic resonance imaging [9]. Therefore, there is a possibility that detecting hip fractures from plain radiographs is also challenging for AI systems.

✉ Takaaki Urakawa
takaaki-u@mwe.biglobe.ne.jp

[1] Department of Orthopedic Surgery, Tsuruoka Municipal Shonai Hospital, 4-20 Izumi-machi, Tsuruoka-shi, Yamagata 997-8515, Japan

[2] Division of Orthopedic Surgery, Department of Regenerative and Transplant Medicine, Niigata University Graduate School of Medical and Dental Sciences, 1-757 Asahimachi-Dori, Niigata 951-8510, Japan

[3] Center for Integrated Human Brain Science, Brain Research Institute, University of Niigata, 1-757 Asahimachi, Niigata 951-8585, Japan

As a first step, we examined the ability of the VGG_16 network to detect intertrochanteric hip fractures, one type of hip fracture, from anterior-view proximal femoral radiographs. The aim of this study was to compare the performance of AI to that of orthopedic surgeons.

Materials and methods

Patient enrolment

We performed a retrospective diagnostic study using our institutional surgical registry database with approval from the institutional review board. We identified all consecutive patients with intertrochanteric hip fractures who were treated using compression hip screws between January 2006 and July 2017. In total, 1773 patients (286 men and 1487 women) were enrolled. The patients' mean age at the time of injury was 85 years (range, 29–104 years).

Data preparation and image selection

Anterior-view hip radiographs were taken with the patient's legs internally rotated. The field of view was 429 × 352 mm. Images of 1773 patients were reviewed by a single board-certified orthopedic surgeon (T.U.) using a Digital Imaging and Communications in Medicine viewer (View R; Yokogawa, Tokyo, Japan) on a color liquid crystal display monitor (FlexScan S2431W; EIZO, Ishikawa, Japan) (resolution, 1920 × 1200, brightness, 450 cd/m^2, contrast ratio, 1000:1). Of 1773 patients, 1626 (91.7%) patients were diagnosed only by anterior-view radiographs. Of the remaining 147 patients, 50 patients were diagnosed using lateral-view radiographs, seven patients were diagnosed using computed tomography with multiplanar reconstruction, and 90 patients were diagnosed by magnetic resonance imaging (coronal T1-weighted image). Furthermore, we compared these diagnosed sides with sides that had been surgically treated, and all image diagnoses were confirmed to coincide.

Anterior-view hip radiographs, which were the same as those used for clinical diagnosis, were subsequently exported from our institutional picture archiving and communication system server using a Digital Imaging and Communications in Medicine viewer (View R; Yokogawa). These radiographs were exported with a matrix size of 1562 × 915 pixels. The Joint Photographic Experts Group format was selected as one of the required formats for further analyses. The image data size was compressed and reduced to 67% of the original size. A board-certified orthopedic surgeon (T.U.) confirmed that 1626 detectable fractures on the original anterior-view hip Digital Imaging and Communications in Medicine images could be also visible on their exported Joint Photographic Experts Group images. Then, the proximal femurs of the fractured and non-fractured sides were cropped with a matrix size of 300 × 300 pixels by T.U. (Fig. 1). This matrix size could include the femoral head and greater and lesser trochanters on the exported whole hip radiographs. We excluded images showing pseudarthrosis after femoral neck fracture (one non-fractured hip) or artificial objects such as prostheses (50 non-fractured hips); screws for multiple pinning (four non-fractured hips); compression hip screws (134 non-fractured hips); Ender nails (one non-fractured hip); intramedullary nails (nine non-fractured hips); and bone cement (one non-fractured hip) in the proximal femoral region. This left us with 3346 hips (1773 fractured and 1573 non-fractured hips) to be used for further analyses. For these analyses, all 3346 hip images were first randomly shuffled and then split into a training set of 2678 images (1408 fractured and 1270 non-fractured images), a validation set of 334 images (185 fractured and 149 non-fractured images), and a test set of 334 images (180 fractured and 154 non-fractured images).

Deep learning framework and model

We used the deep learning framework of TensorFlow, an open-source software library for machine learning [10]. We downloaded the VGG_16 pre-trained model [5] from TensorFlow's official site [11]. The model was released under the Creative Commons Attribution License. We used it for transfer learning purpose because pre-trained models reduce the training time and the number of required images to make an appropriate classifier [12].

Training operation

Because the CNN has many parameters, the network has an excessive amount of freedom. It can easily fit complex training data, decreasing generalizability [13]. Therefore, we used three regularization techniques to avoid overfitting to the training set [13]. One was data augmentation, which generates new training instances from existing ones. It was used to artificially boost the size of the training set. TensorFlow provides an application programming interface of ImageDataGenerator for this purpose [14]. The ImageDataGenerator randomly selected 50 images from the training set at every iteration. Then, each selected image was modified to the rotation angle range of 5.0°, width shift range of 0.2, height shift range of 0.2, shear range of 0.1, zoom range of 0.2, and horizontal flip in 50%. The second was L2 regularization, which made the network robust against outliers in the training data (weight decay, 0.001). The third was early stopping, which adopts the best performance parameters on the validation set (not on the training set) as final network parameters. In total, 2650 iterations, or training of 132,500 (2650 × 50) augmented images, were performed using the adaptive moment estimation (Adam) optimizer [15]. It tweaks the network parameters iteratively in

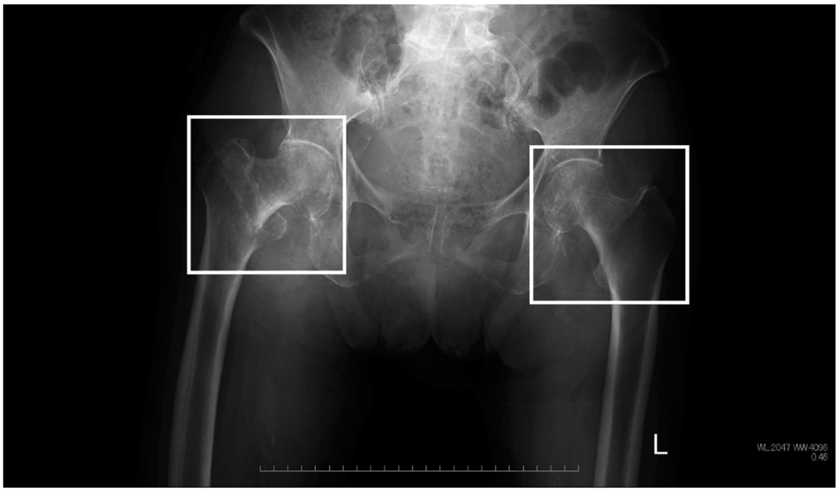

Fig. 1 Proximal femurs are cropped from an anterior-view hip radiograph. Each image is labeled as either a fractured or a non-fractured hip

order to enhance the network performance. To speed up training, exponential learning rate scheduling was applied with the following parameters: initial learning rate, 0.0001; decay steps, 265 iterations; and decay rate, 0.8.

Classification of test set images by the VGG_16 network

The final network parameters were restored, and each image in the test set was classified as either fractured or non-fractured.

Orthopedic surgeons' diagnostic performance

Five orthopedic surgeons, including two board-certified (S.G. and K.K.) and three non-board-certified surgeons (Y.T., Y.F., and J.W.), reviewed the same 334 test set images (cropped proximal femoral images) at the same resolution as the network on a color liquid crystal display monitor (FlexScan S2431W; EIZO) (resolution, 1920 × 1200, brightness, 450 cd/m^2, contrast ratio, 1000:1) and classified each image as either a fractured or non-fractured hip. They routinely adjusted the brightness, contrast, and zoom settings when fractures were unclear or non-detectable with default contrast. In cases where the surgeons did not agree whether hips were fractured or non-fractured, final diagnoses were decided by a majority vote.

Comparison of performances between the VGG_16 network and orthopedic surgeons

First, accuracies, sensitivities, specificities, and areas under the receiver operating characteristic curves were calculated for the VGG_16 network and for the orthopedic surgeons. Then, distributions for these outcomes were computed using bootstrapping with 10,000 bootstraps, and 95% CIs were calculated. Finally, *t* tests were conducted to compare these distributions. All statistical analyses were performed using SciPy, an open-source scientific tool for Python, and significance was set at $p < 0.05$.

Results

The learning processes of the network in the training and validation sets are shown in Fig. 2a, b. The network showed the best performance at 1457 iterations; therefore, the weight

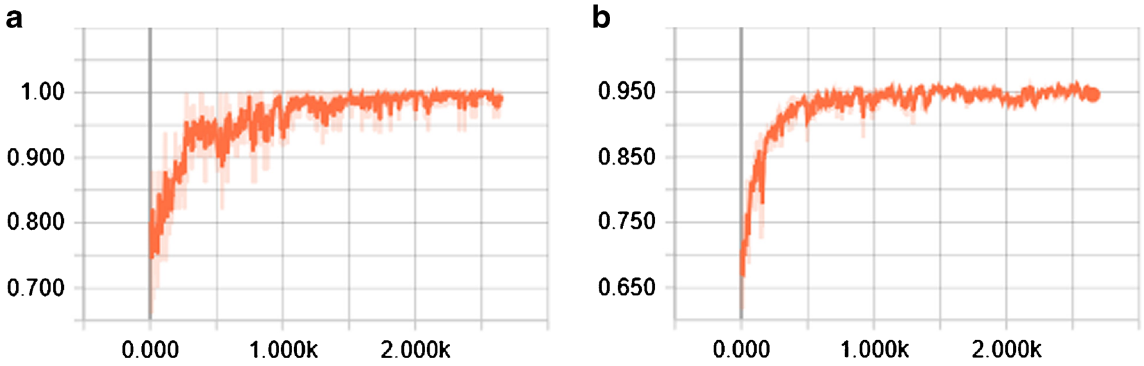

Fig. 2 The learning processes of the Visual Geometry Group 16-layer (VGG_16) network in the training (**a**) and validation (**b**) sets. The VGG_16 network performed at an accuracy of about 95.0% in the validation set. The *horizontal axis* represents iterations, and the *vertical axis* shows accuracy scores

Table 1 Comparison of performances between the VGG_16 network and orthopedic surgeons

	VGG_16 network	Orthopedic surgeons	*p* value
Accuracy, (%) (95% CI)	319/334 (95.5) (93.1–97.6)	308/334 (92.2) (89.2–94.9)	< 0.001
Sensitivity, (%) (95% CI)	169/180 (93.9) (90.1–97.1)	159/180 (88.3) (83.3–92.8)	< 0.001
Specificity, (%) (95% CI)	150/154 (97.4) (94.5–99.4)	149/154 (96.8) (95.1–98.4)	< 0.001
AUC (95% CI)	0.984 (0.970–0.996)	0.969 (0.951–0.984)	< 0.001

VGG_16 Visual Geometry Group 16-layer, *CI* confidence interval, *AUC* area under the curve

and bias parameters at this point were used for testing the general network performance. In the test set, the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve of the VGG_16 network exceeded those of the orthopedic surgeons (Tables 1 and 2 and Fig. 3). Eleven and four images for the VGG_16 network and 21 and five images for orthopedic surgeons were false negatives and false positives, respectively. Receiver operating characteristic curves for the VGG_16 network and the orthopedic surgeons' diagnostic performance are plotted in Fig. 4. For reference, representative fractured hip images are presented in Fig. 5a, b.

Discussion

Our study demonstrated that the performance of the VGG_16 network (95.5% accuracy) exceeded that of orthopedic surgeons (92.2% accuracy) in detecting intertrochanteric fractures from proximal femoral images (cropped images from whole hip radiographs). Previous studies have also investigated the feasibility of using AI for fracture detection [3, 6, 7]. One key study was performed by Olczak et al., in which the VGG_16 network was used to identify fractures from whole hand, wrist, and ankle radiographs with an accuracy of 83% [3]. In contrast, radiologists in that study performed at an accuracy of 82%. Although the exact reasons why our VGG_16 network performance exceeded that in the study by Olczak et al. are unclear, one possible reason is the difference between the analyzed extremity regions of the radiographs. Another possible reason is that we had the network

Table 2 Diagnostic performances of the five orthopedic surgeons

	Accuracy, (%) (95% CI)	Sensitivity, (%) (95% CI)	Specificity, (%) (95% CI)
Reviewer S.G.	306/334 (91.6) (88.6–94.3)	166/180 (92.2) (88.1–95.8)	140/154 (90.9) (86.2–95.2)
Reviewer K.K.	291/334 (87.1) (83.2–90.7)	139/180 (77.2) (70.9–83.2)	152/154 (98.7) (96.6–100.0)
Reviewer Y.T.	292/334 (87.4) (83.8–90.7)	171/180 (95.0) (91.5–97.8)	121/154 (78.6) (71.8–84.6)
Reviewer Y.F.	304/334 (91.0) (87.7–94.0)	160/180 (88.9) (83.9–93.2)	144/154 (93.5) (89.3–97.2)
Reviewer J.W.	286/334 (85.6) (81.7–89.2)	140/180 (77.8) (71.6–83.7)	146/154 (94.8) (91.0–98.1)

CI confidence interval

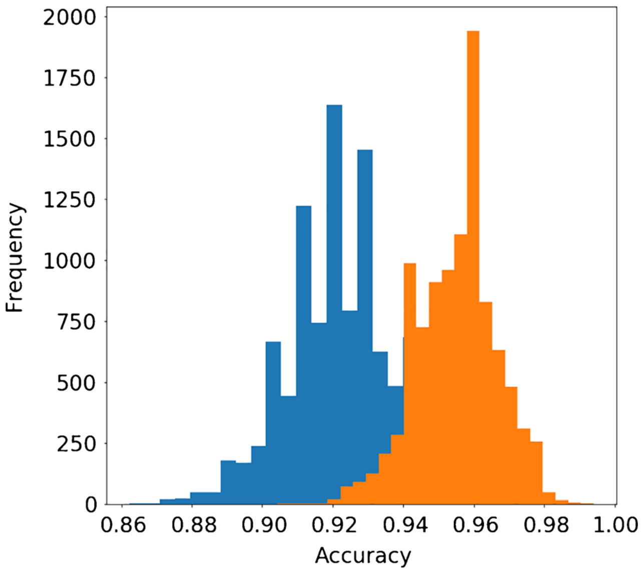

Fig. 3 Histograms of the accuracy scores computed using bootstrapping with 10,000 bootstraps. *Red and blue bins* represent histograms of the Visual Geometry Group 16-layer network and orthopedic surgeons, respectively

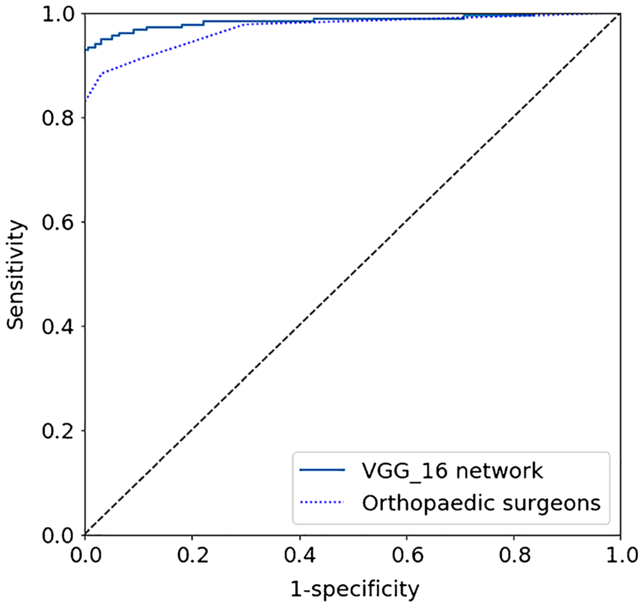

Fig. 4 The receiver operating characteristic curves of the Visual Geometry Group 16-layer (VGG_16) network and orthopedic surgeons. The areas under the curves are 0.984 and 0.969 for the VGG_16 network and orthopedic surgeons, respectively

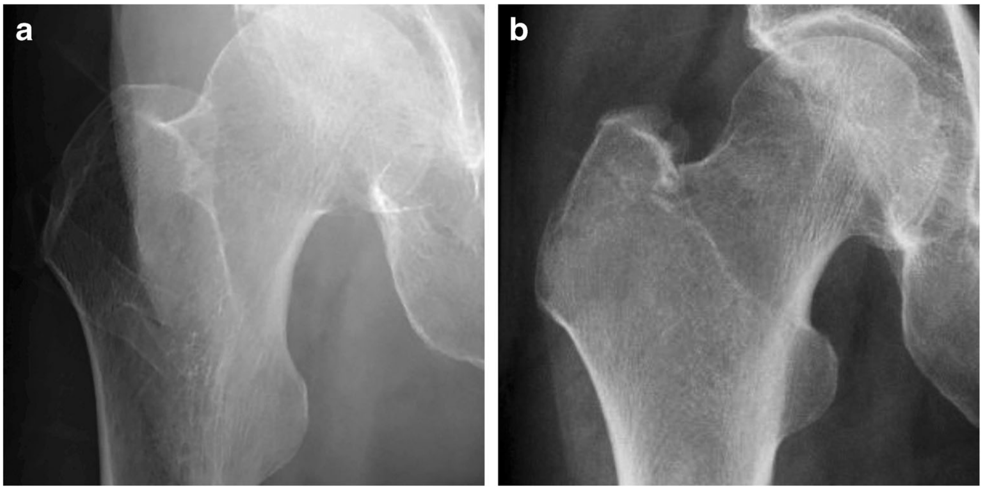

Fig. 5 Representative fractured hip images. **a** Example of an image from which both the network and orthopedic surgeons could not detect fractures. **b** Example of an image from which the network could detect fracture but orthopedic surgeons had some difficulty in recognizing the image as a fractured hip (detectable rate of 40%, two out of five orthopedic surgeons)

analyze fractures from cropped radiographs rather than from whole hip radiographs. Chung et al. performed a study of proximal humerus fracture detection using manually cropped anterior-view shoulder radiographs [7]. They achieved superior network accuracy (96%) compared to that of orthopedic surgeons (93%). There is a possibility that using cropped proximal femoral radiographs contributed to improvement of the network performance (95.5% accuracy) because the network did not need to analyze irrelevant regions, such as the pelvic bone or femoral shaft.

In the process of training a deep neural network to detect intertrochanteric fractures on radiographs, we had to obtain many images of intertrochanteric fractured and non-fractured (control). We thought proximal femoral images without showing pseudarthrosis after femoral neck fracture or artificial objects were more appropriate as controls against images with intertrochanteric fracture. Therefore, we excluded 200 images showing pseudarthrosis after femoral neck fracture or artificial objects from 1773 non-fractured hip images. According to a survey by Dinah [16], 11.8% of patients with hip fractures had undergone previous surgery for a contralateral hip fracture. Therefore, an exclusion rate of 11.2% (200/1773) from non-fractured hip images seems to be an appropriate rate.

We cropped fractured and non-fractured proximal femurs from the whole hip radiograph of each patient. Although manual cropping was performed by a single orthopedic surgeon (T.U.), there is a possibility that bias was introduced through this process. Automated proximal femur detection and cropping by antecedent neural network can be a solution to decrease this bias.

A lateral-view radiograph of a fractured hip is taken by flexing the contralateral hip and knee to 90° and aiming the beam into the groin, parallel to the floor and perpendicular to the femoral neck [17]. Conversely, obtaining a lateral-view radiograph of a non-fractured hip is difficult because flexing the hip and knee of the fractured leg to 90° is too painful for the patient. Therefore, lateral-view radiographs of non-fractured hips generally do not exist in typical clinical practice. As this study was a retrospective study, we could not obtain controls against the lateral-view radiographs of fractured hips or train the network using lateral-view images. If the study had been performed prospectively, the lateral-view radiograph of the non-fractured side could have been taken after the patient's fractured hip pain relieved. The ensemble of AI analyses using anterior-view and lateral-view radiographs has the potential to enhance the detection rate of intertrochanteric fractures.

Each image analyzed by the VGG_16 network included the femoral intertrochanteric region and the head and neck region. The fractures in these regions cannot be distinguished by clinical features; therefore, it is not realistic to limit the field of analysis as we have done. To make a claim that the VGG_16 network is feasible, further studies are needed to determine whether a deep CNN can discriminate these fractures by assembling three region-specific binary classifiers (classifier of the femoral head region, that of the neck region, and that of the intertrochanteric region) or by making a multiclass classifier of these three regions.

To summarize, we performed this study to compare the performances in diagnosing intertrochanteric hip fractures from proximal femoral radiographs between a deep CNN and orthopedic surgeons. The network achieved an accuracy of 95.5%, which exceeded that of orthopedic surgeons under restricted or limited conditions. Therefore, deep CNN has a significant potential to be a useful tool for screening of fractures from plain radiographs, especially in emergency rooms, where orthopedic surgeons are not readily available.

Acknowledgements The authors express their appreciation to Kamimura K, Fujita Y, and Wakui J for reviewing images and would like to thank Editage (www.editage.jp) for English language editing.

Funding The study was conducted without any external funding or grant.

Compliance with ethical standards

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required.

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. Med Image Anal. 2017;41:63–73.
2. Jamaludin A, Lootus M, Kadir T, et al. Genodisc consortium. ISSLS prize in bioengineering science 2017: automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. Eur Spine J. 2017;26(5):1374–83.
3. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiography. Acta Orthop. 2017;88(6):581–6.
4. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging. 2017;30(4):427–41.
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [Internet]. 2014 [cited 2017 Dec 10] Available from: https://arxiv.org/abs/1409.1556
6. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73(5):439–45.
7. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 2018; https://doi.org/10.1080/17453674.2018.1453714.
8. Hagino H, Endo N, Harada A, et al. Survey of hip fractures in Japan: recent trends in prevalence and treatment. J Orthop Sci. 2017;22(5):909–14.
9. Pejic A, Hansson S, Rogmark C. Magnetic resonance imaging for verifying hip fracture diagnosis why, when and how? Injury. 2017;48(3):687–91.
10. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv [Internet]. 2016 [cited 2017 Dec 10] Available from: https://arxiv.org/abs/1603.04467
11. No authors listed. TesnorFlow-Slim image classification model library. [Internet]. [cited 2017 Dec 10] Available from: https://github.com/tensorflow/models/tree/master/research/slim
12. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35(5):1285–98.
13. Geron A. Training deep neural network. In: Geron A, editor. Hands-on machine learning with scikit-learn & TensorFlow. Sebastopol: O'Reilly Media; 2017. p. 275–312.
14. No authors listed. tf.keras.preprocessing.image.ImageDataGenerator. [Internet]. [cited 2017 Dec 10] Available from: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator
15. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv [Internet]. 2014 [cited 2017 Dec 10] Available from: https://arxiv.org/abs/1412.6980
16. Dinah AF. Sequential hip fractures in elderly patients. Injury. 2002;33(5):393–4.
17. Baumgaertner MR, Higgins TF. Femoral neck fractures. In: Bucholz BW, Heckman JD, editors. Rockwood and Green's fractures in adults. 5th ed. Philadelphia: Lippincott Williams & Wilkins; 2001. p. 1579–634.